Consistance sous un modèle de réponse de la fonction de répartition estimée en présence de données manquantes

Hélène BOISTARD (*), Guillaume CHAUVET(**), David HAZIZA(***)

(*)Université Toulouse 1
(**) Crest, Ensai
(***) Université de Montréal

Introduction

On rencontre des problèmes de données manquantes dans les enquêtes quand certaines des unités refusent de répondre, ou quand il est impossible de les contacter. Dans le contexte des Essais Cliniques, on peut également obtenir des données manquantes pour les sujets qui abandonnent ou sont perdus de vue pendant le traitement. Les estimateurs non ajustés pour la non-réponse peuvent être fortement biaisés si les répondants diffèrent des non-répondants au regard des variables étudiées. Il est donc souhaitable de développer des procédures d'estimation conduisant à des biais aussi faibles que possible. Les procédures d'estimation doublement robustes ont été largement étudiées dans le cadre de la Statistique classique, voir par exemple [10], [12], [13], [14], [15], [7], [2]. Dans le cadre d'un sondage en population finie, l'estimation doublement robuste a été étudiée par [9], [8] et [6].

Nous nous intéressons ici au cas où une méthode d'imputation simple est utilisée pour remplacer une valeur manquante par une valeur artificielle. L'objectif principal de l'imputation est de réduire le biais d'estimation, ce qui nécessite de disposer de variables auxiliaires bien explicatives et disponibles pour toutes les unités de l'échantillon. Les méthodes d'imputation simples sont souvent utilisées dans les agences de Statistique pour le traitement de la non-réponse partielle. Afin d'étudier les propriétés des estimateurs, nous considérons deux approches : (i) l'approche par le modèle de non-réponse (NM), et (ii) l'approche par le modèle d'imputation (IM) qui nécessite la spécification d'un modèle décrivant la distribution de la variable étudiée. Un estimateur est dit doublement robuste s'il reste asymptotiquement sans biais et consistant quand l'un des deux modèles spécifiés (NM ou IM) est correct. Les procédures doublement robustes offrent, dans une certaine mesure, une protection contre une mauvaise spécification de l'un des deux modèles. Cependant, la littérature mentionnée ci-dessus traite seulement le cas où l'on estime la valeur moyenne de la variable d'intérêt. A notre connaissance, le cas d'un estimateur doublement robuste de la fonction de répartition est peu (ou pas) abordé dans la littérature.

Dans ce travail, nous considérons le cas d'un mécanisme d'imputation des valeurs manquantes par la régression aléatoire pondérée (motivé par le modèle IM). Plus précisément, pour chaque non-répondant, un résidu observé pour un répondant est sélectionné au hasard avec une probabilité proportionnelle à un poids d'imputation, et utilisé dans l'imputation de la valeur manquante. Nous étudions les propriétés de double robustesse de la fonction de répartition estimée, en fonction des poids d'imputation utilisés. Les propriétés de la fonction de répartition estimée sont évaluées théoriquement, et à l'aide d'une étude par simulations.

1 Notations

On considère une population finie U de taille N. Nous nous intéressons à l'estimation de la fonction de répartition $F_{N,y}(t) = N^{-1}\sum_{i\in U} 1(y_i \leq t)$, où y désigne une variable d'intérêt et $1(.)$ désigne la fonction indicatrice usuelle. Un échantillon s, de taille n, est sélectionné selon un plan de sondage $p(\cdot)$. Soit $d_i = 1/\pi_i$, le poids de sondage de l'unité i, où π_i désigne la probabilité d'inclusion d'ordre 1 dans l'échantillon. En l'absence de non-réponse, un estimateur basé sur les données complètes est donné par l'estimateur par expansion

$$\hat{F}_{N,y}(t) \;=\; \sum_{i\in s} \tilde{d}_i 1(y_i \leq t), \tag{1}$$

avec $\tilde{d}_i = \left(\sum_{j\in s} d_j\right)^{-1} d_i$.

Quand certaines des valeurs de la variable d'intérêt y sont manquantes, un estimateur de $F_{N,y}(t)$ est donné par l'estimateur imputé

$$\hat{F}_{I,y}(t) \;=\; \sum_{i\in s} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i\in s} \tilde{d}_i (1-r_i) 1(y_i^* \leq t), \tag{2}$$

avec y_i^* la valeur imputée pour remplacer une valeur manquante y_i, et r_i l'indicatrice de réponse pour l'unité i qui vaut 1 si y_i est observée et 0 si y_i est manquante. On note également s_r et s_m les sous-échantillons de répondants et de non-répondants, respectivement.

Nous supposons qu'un vecteur de variables auxiliaires $\mathbf{x}$ est disponible pour toutes les unités échantillonnées (répondantes ou non). Nous postulons le modèle suivant pour la distribution conditionnelle de y_i sachant $\mathbf{x}_i$:

$$m: \; y_i = m(\mathbf{x}_i; \boldsymbol{\beta}_0) + \sigma\sqrt{v_i}\epsilon_i, \tag{3}$$

où $m(\mathbf{x}_i, \boldsymbol{\beta})$ désigne une fonction de $\boldsymbol{\beta}$ continument différentiable, σ^2 est un paramètre inconnu et v_i est une constante connue. Les résidus ϵ_i sont des variables aléatoires indépendantes et identiquement distribuées de moyenne 0 et de variance 1. Le modèle (3) est appelé le modèle d'imputation. Dans ce qui suit, nous considérons le modèle de régression linéaire avec $m(\mathbf{x}_i; \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$.

En utilisant l'équation (3), il peut être tentant d'estimer $F_{N,y}(t)$ par

$$\hat{F}_{I,y}(t) \;=\; \sum_{i\in s} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i\in s} \tilde{d}_i (1-r_i) 1\{m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \leq t\}, \tag{4}$$

où $\hat{\boldsymbol{\beta}}$ désigne un estimateur consistant du vrai paramètre $\boldsymbol{\beta}_0$. Souvent, un estimateur consistant $\hat{\boldsymbol{\beta}}$ est obtenu en résolvant

$$\sum_{i \in s} d_i r_i \left\{ y_i - m\left(\mathbf{x}_i; \boldsymbol{\beta}\right) \right\} \mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta}) \quad = \quad \mathbf{0}, \tag{5}$$

pour une fonction $\mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta})$, ce qui se justifie sous l'hypothèse de données manquantes au hasard ou Missing At Random (MAR). Cette hypothèse peut s'exprimer sous la forme

$$E\left(y_i \mid \mathbf{x}_i, r_i = 1\right) = E\left(y_i \mid \mathbf{x}_i, r_i = 0\right). \tag{6}$$

Notons que, sous certaines conditions de régularité, la solution $\hat{\boldsymbol{\beta}}$ à l'équation (5) est consistante pour $\boldsymbol{\beta}_0$ si le modèle (3) et la condition MAR (6) sont vérifiés. Cependant, l'estimateur (4) est généralement biaisé. Pour résoudre ce problème, [3] ont proposé un estimateur corrigé du biais.

Une autre approche pour estimer $F_{N,y}(t)$ consiste à utiliser une méthode d'imputation aléatoire. Une valeur manquante y_i est remplacée par

$$y_i^* \quad = \quad \mathbf{x}_i^\top \, \hat{\mathbf{B}}_r + \hat{\sigma} \sqrt{v_i} \epsilon_i^* \text{ pour } i \in s_m, \tag{7}$$

où $\hat{\sigma}$ est un estimateur de σ et

$$\hat{\mathbf{B}}_r \quad = \quad \left[\sum_{i \in s} \omega_i v_i^{-1} r_i \mathbf{x}_i \mathbf{x}_i^\top \right]^{-1} \sum_{i \in s} \omega_i v_i^{-1} r_i \mathbf{x}_i y_i, \tag{8}$$

avec ω_i un poids d'imputation attaché à l'unité i. Bien qu'elles puissent être générées selon une distribution paramétrique donnée, il est naturel de sélectionner les quantités ϵ_i^* au hasard parmi les résidus observés sur les répondants. Plus précisément, les résidus aléatoires ϵ_i^* sont sélectionnés indépendamment et avec remise dans l'ensemble $E_r = \{\tilde{e}_j; j \in s_r\}$, des résidus standardisés observés sur les unités répondantes, avec des probabilités

$$pr\left(\epsilon_i^* = \tilde{e}_j\right) \quad = \quad \tilde{\omega}_j, \tag{9}$$

où

$$\begin{aligned} \tilde{\omega}_j \quad &= \quad \left(\sum_{l \in s} \omega_l r_l\right)^{-1} \omega_j, \\ \tilde{e}_j \quad &= \quad e_j - \bar{e}_r, \end{aligned}$$

avec

$$\begin{aligned} e_j \quad &= \quad \hat{\sigma}^{-1} v_j^{-1/2} \left\{ y_j - \mathbf{x}_j^\top \, \hat{\mathbf{B}}_r \right\}, \\ \bar{e}_r \quad &= \quad \sum_{j \in s} \tilde{\omega}_j r_j e_j. \end{aligned}$$

Nous supposons qu'il existe un vecteur μ de constantes connues, tel que $v_i^{1/2} = \mu^\top \mathbf{x}_i$, de sorte que $\bar{e}_r = 0$.

2 Estimation doublement robuste

Dans cet article, les propriétés des estimateurs sont étudiées sous deux approches distinctes : (i) l'approche par le modèle de non-réponse (NM) et (ii) l'approche par le modèle d'imputation (IM). Ces deux approches sont décrites ci-dessous :

(i) l'approche NM : nous faisons des hypothèses explicites (appelées le modèle de non-réponse) sur le mécanisme (inconnu) de non-réponse. L'inférence est réalisée par rapport à la distribution de probabilité induite par le plan de sondage et le modèle (postulé) de non-réponse. Nous supposons que les unités échantillonnées répondent indépendamment les unes des autres. Nous supposons également que la probabilité de réponse à la variable y, notée $p_i = \Pr(r_i = 1)$, suit le modèle de régression logistique

$$p_i = p_i(\boldsymbol{\phi}_0) \quad = \quad \frac{\exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_i\right)}{1 + \exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_i\right)} \tag{10}$$

pour un certain $\boldsymbol{\phi}_0$, et que la variable constante est incluse dans les $\mathbf{x}_i$. Le modèle (10) est appelé le modèle de non-réponse. Il est important de noter que sous l'approche NM, aucune hypothèse n'est faite sur la distribution de la variable d'intérêt y. En d'autres termes, les propriétés des estimateurs sous cette approche ne dépendent pas de la validité du modèle (3).

(ii) l'approche IM : nous faisons des hypothèses explicites sur la distribution de la variable d'intérêt, sous la forme du modèle d'imputation (3). Contrairement à l'approche NM, le mécanisme de réponse n'est pas explicitement spécifié, même si nous faisons l'hypothèse qu'il est non-confondu ; voir par exemple [11].

Nous considérons ici une procédure d'imputation doublement robuste (sous certaines hypothèses) pour la fonction de répartition, i.e. conduisant à une estimation consistante de la fonction de répartition sous chacune des deux approches (NM ou IM). Nous supposons qu'une valeur manquante y_i est imputée selon l'équation (7), avec un poids d'imputation donné par $\omega_j = d_j(1 - p_j)/p_j$, voir [6]. Notons qu'en pratique, la probabilité de réponse p_j est inconnue et doit être estimée.

Nous introduisons ici une version déterministe du modèle d'imputation, qui nous sera utile dans la suite. On note :

$$y_i \quad = \quad \mathbf{x}_i^\top \, \mathbf{B}_U + \sigma_U \, \sqrt{v_i} \, E_i, \tag{11}$$

avec

$$\mathbf{B}_U \quad = \quad \left\{ \sum_{i \in U} (1 - p_i) v_i^{-1} \mathbf{x}_i \mathbf{x}_i^\top \right\}^{-1} \sum_{i \in U} (1 - p_i) v_i^{-1} \mathbf{x}_i y_i, \tag{12}$$

$$\sigma_U^2 \quad = \quad \left\{ \sum_{i \in U} (1 - p_i) \right\}^{-1} \sum_{i \in U} (1 - p_i) v_i^{-1} \left(y_i - \mathbf{x}_i^\top B_U \right)^2. \tag{13}$$

On note également $F_{N,E}(\cdot)$ et $\hat{F}_{N,E}(\cdot)$ la fonction de répartition associée à la variable de résidus E_i, et son estimateur par expansion, obtenues à partir de $F_{N,y}(\cdot)$ et $\hat{F}_{N,y}(\cdot)$, respectivement, en remplaçant la variable y_i par la variable E_i.

Les valeurs imputées y_i^*, définies par (7), peuvent être obtenues de la façon suivante. Tout d'abord, pour chaque unité $i \in s_m$, on sélectionne indépendamment des résidus

aléatoires $\hat{E}_i$ dans l'ensemble $G_r = \{E_j; j \in s_r\}$ des résidus exacts définis par (11). On note $j(i)$ le donneur sélectionné pour l'unité i et

$$\hat{y}_i \;=\; \mathbf{x}_i^\top \mathbf{B}_U + \sigma_U \sqrt{v_i} \hat{E}_i.$$

Dans $\hat{y}_i$ les paramètres $\mathbf{B}_U$ et σ_U sont inconnus, et sont remplacés par leurs estimateurs $\hat{\mathbf{B}}_r$, $\hat{\sigma}$. Le résidu exact $E_{j(i)}$ est également inconnu, et remplacé par le résidu estimé $\tilde{e}_{j(i)}$, pour obtenir la valeur imputée finale y_i^*.

3 Principaux résultats

Nous étudions maintenant les propriétés asymptotiques de la fonction de répartition estimée sous la méthode d'imputation aléatoire étudiée. Nous supposons qu'il existe une suite de plans de sondage et de populations finies, indexées par ν, telles que la taille de la population N_ν, la taille d'échantillon n_ν et le nombre de répondants $n_{r\nu}$ tendent vers l'infini quand $\nu \to \infty$. Bien que nous supprimions l'indice ν pour ne pas alourdir les notations, les limites sont comprises comme étant en $\nu \to \infty$.

Sous de faibles hypothèses de régularité, le théorème 1 de [4] implique que sous l'approche IM, $\hat{F}_{I,y}(t) - F_{N,y}(t)$ tend en probabilité vers 0. Il suffit donc de montrer le résultat de consistance sous l'approche NM. Nous faisons les hypothèses de régularité suivantes :

C1a : Pour tout $i \neq j \in U$, $\pi_{ij} - \pi_i\, \pi_j \leq 0$;

C1b : $\max_{i \neq j \in U} |\pi_{ij} - \pi_i\, \pi_j| = O(n^{-1})$;

C2 : Il existe une constante $0 < f < 1$ telle que $n/N \to f$;

C3 : $\max\ N^{-1}\, d_i = O(n^{-1})$;

C4 : Il existe une constante $0 < \kappa < 1$ telle que $\kappa < p_i$ pour tout $i \in s$;

C5 : $F_{N,E}(\cdot)$ est uniformément convergent, au sens où $\forall \epsilon > 0\ \exists \eta > 0$ tel que $|t - u| \leq \eta \Rightarrow |F_{N,E}(t) - F_{N,E}(u)| \leq \epsilon$.

C6 : Les composantes du vecteur de variables auxiliaires $\mathbf{x}_i$ ainsi que le nombre K de variables auxiliaires sont bornés.

C7 : $\hat{\mathbf{B}}_r \to_{\mathbb{P}} \mathbf{B}_U$, où $\to_{\mathbb{P}}$ désigne la convergence en probabilité ;

Les conditions C1a, C1b, C2 et C3 sont des hypothèses de régularité standard, voir par exemple Breidt et Opsomer (2000). La condition C3 garantit qu'aucun poids extrême ne domine les autres. La condition C4 assure que les probabilités de réponses admettent une borne inférieure strictement positive. La condition C7 assure que $\hat{\mathbf{B}}_r$ est un estimateur convergent de B_U.

Avant de démontrer le résultat principal, nous avons besoin de lemmes intermédiaires.

Lemme 1. *Supposons que la condition C1a ou C1b est vérifiée, et que les conditions C2–C7 sont vérifiées. Soit*

$$T_2 \;=\; \sum_{i \in s} \tilde{d}_i (1 - r_i) \left\{ 1(y_i^* \leq t) - 1(\hat{y}_i \leq t) \right\}. \tag{14}$$

Alors T_2 tend en probabilité vers 0.

Démonstration. Donnée en annexe 5. □

Dans ce qui suit, nous nous restreignons au cas du hot-deck aléatoire, pour lequel une valeur manquante y_i est remplacée en sélectionnant au hasard et avec remise un donneur $j \in s_r$, et en remplaçant la valeur manquante par y_j. Le hot-deck est un cas particulier de la méthode d'imputation étudiée, obtenu en prenant $\mathbf{x}_i = x_i = 1$, et $v_i = 1$. Plus précisément, les valeurs imputées sont données par

$$y_i^* \quad = \quad y_{(j)} \text{ pour } i \in s_m, \tag{15}$$

où $y_{(j)}$ désigne une valeur tirée au hasard et avec remise dans l'ensemble des valeurs y_j observées sur les unités répondantes, avec des probabilités

$$pr\left(y_i^* = y_j\right) \quad = \quad \tilde{\omega}_j. \tag{16}$$

Les résultats obtenus ci-dessous s'étendent facilement au cas du hot-deck dans des classes d'imputation (voir par exemple [5]).

Lemme 2. *Supposons que la condition C1a ou C1b est vérifiée, et que les conditions C2–C7 sont vérifiées. Soit*

$$T_1 \quad = \quad \sum_{i \in s} \tilde{d}_i (1 - r_i) \left\{ 1(\hat{y}_i \leq t) - 1(y_i \leq t) \right\}. \tag{17}$$

Alors dans le cas du hot-deck aléatoire, T_1 tend en probabilité vers 0.

Démonstration. Donnée en annexe 5. □

Nous pouvons maintenant énoncer notre résultat principal.

Théorème 1. *Supposons que la condition C1a ou C1b est vérifiée, et que les conditions C2–C7 sont vérifiées. Alors le hot-deck aléatoire donne une estimation consistante de la fonction de répartition sous l'approche NM, i.e. $\hat{F}_{I,y}(t) - F_{N,y}(t)$ tend en probabilité vers 0.*

Démonstration. Notons tout d'abord que l'erreur totale de $\hat{F}_{I,y}(t)$ peut s'écrire sous la forme

$$\hat{F}_{I,y}(t) - F_{N,y}(t) \quad = \quad \left\{ \hat{F}_{I,y}(t) - \hat{F}_{N,y}(t) \right\} + \left\{ \hat{F}_{N,y}(t) - F_{N,y}(t) \right\}.$$

Sous des conditions standard de régularité (voir par exemple Isaki et Fuller, 1981), nous avons $\hat{F}_{N,y}(t) - F_{N,y}(t) = O_p(n^{-1/2})$. Il suffit donc de montrer que $\hat{F}_{I,y}(t) - \hat{F}_{N,y}(t) \rightarrow_{\mathbb{P}} 0$.

Le terme d'erreur peut se décomposer sous la forme

$$\hat{F}_{I,y}(t) - \hat{F}_{N,y}(t) = T_1 + T_2. \tag{18}$$

Le résultat s'obtient donc par application des lemmes 1 et 2. □

4 Etude par simulations

Nous avons réalisé une petite étude par simulations afin de tester les performances de différentes méthodes d'imputation, en termes de biais relatif et d'efficacité relative. Nous avons tout d'abord généré une population finie de taille $N = 10,000$, contenant une variable d'intérêt y et deux variables auxiliaires x_1 et x_2. Les variables auxiliaires

ont été générées selon une distribution Gamma de paramètres d'échelle 2 et de paramètre d'intensité 5. Sachant les valeurs de x_1 et de x_2, les valeurs de y ont été générées selon le modèle

$$y_i = \beta_0 + \beta_1 \; x_{1i} + \beta_2 \; x_{2i} + \eta_i. \tag{19}$$

Les paramètres β_0, β_1 et β_2 ont été fixés à 10, 1 et 1, respectivement. Les η_i ont été générés selon une distribution normale de moyenne 0 et de variance σ^2, dont la valeur a été choisie pour obtenir un coefficient de détermination (R^2) approximativement égal à 0.7.

Nous nous intéressons à l'estimation de la fonction de répartition $F_{N,y}(t)$ pour $t = t_\alpha$, avec t_α le $\alpha^{\text{ème}}$ quantile. Nous avons considéré $\alpha = 0.05, 0.25, 0.50, 0.75$ et 0.95 dans la simulation. Dans la population, nous sélectionnons $1,000$ échantillons de taille $n = 500$ par sondage aléatoire simple sans remise. Puis, dans chaque échantillon sélectionné, la non-réponse est générée selon le mécanisme de réponse suivant :

$$Pr(r_i = 1|x_{1i}, x_{2i}) \;\; = \;\; \frac{\exp\left(-1 + 1.6 \; x_{1i} + 1.6 \; x_{2i}\right)}{1 + \exp\left(-1 + 1.6 \; x_{1i} + 1.6 \; x_{2i}\right)}. \tag{20}$$

La probabilité de réponse moyenne est approximativement égale à 0.6. Dans chaque sous-ensemble de répondants, les probabilités de réponse sont estimées à l'aide d'une régression logistique. En utilisant les covariables $\mathbf{x}_i = (1, x_{1i}, x_{2i})$, cela conduit aux probabilités de réponse estimées $\hat{p}_{1i}$. Afin d'évaluer (dans une certaine mesure) les performances des méthodes d'imputation quand le modèle de réponse est mal spécifié, nous avons utilisé une seconde modélisation des probabilités de réponse utilisant les covariables $\mathbf{x}_i = (1, x_{1i})$ uniquement. Cela conduit aux probabilités de réponse estimées $\hat{p}_{2i}$.

Les valeurs manquantes ont tout d'abord été imputées en utilisant une imputation par la régression aléatoire, avec un modèle d'imputation correctement spécifié. Plus précisément, les valeurs ont été imputées selon le mécanisme d'imputation (7) avec $\mathbf{x}_i = (1, x_{1i}, x_{2i})$ et $v_i = 1$. Le choix $\omega_i = 1$ conduit à une imputation par la régression aléatoire non pondérée (REGI). Le choix $\omega_i = d_i(1-\hat{p}_{1i})\hat{p}_{1i}^{-1}$ conduit à une imputation par la régression aléatoire pondérée avec modélisation correcte de la non-réponse (REGI-P1). Le choix $\omega_i = d_i(1-\hat{p}_{2i})\hat{p}_{2i}^{-1}$ conduit à une imputation par la régression aléatoire pondérée avec modélisation incorrecte de la non-réponse (REGI-P2). Afin d'évaluer les performances des méthodes d'imputation par la régression aléatoire quand le modèle d'imputation est spécifié, nous avons également mis en oeuvre ces trois méthodes avec $\mathbf{x}_i = (1, x_{1i})$ dans le mécanisme d'imputation (7). Les valeurs manquantes ont également été imputées par hot-deck selon le mécanisme d'imputation (15). Le choix $\omega_i = 1$ conduit à une imputation par la hot-deck non pondéré (RHDI). Le choix $\omega_i = d_i(1-\hat{p}_{1i})\hat{p}_{1i}^{-1}$ conduit à une imputation par hot-deck avec modélisation correcte de la non-réponse (RHDI-P1). Le choix $\omega_i = d_i(1-\hat{p}_{1i})\hat{p}_{1i}^{-1}$ conduit à une imputation par hot-deck avec modélisation incorrecte de la non-réponse (REGI-P2).

Pour chacune de ces 9 méthodes d'imputation, nous avons calculé l'estimateur imputé de $F_{N,y}(t)$, $\hat{F}_{I,y}(t)$, donné par (2). Comme mesure du biais de $\hat{F}_{I,y}(t)$, nous avons utilisé le biais relatif (en pourcentage) de Monte Carlo

$$\text{RB}\{\hat{F}_{I,y}(t)\} = \frac{E_{MC}\{\hat{F}_{I,y}(t)\} - F_{N,y}(t)}{F_{N,y}(t)} \times 100, \tag{21}$$

avec $E_{MC}\{\hat{F}_{I,y}(t)\} = \sum_{r=1}^{1000} \hat{F}_{I,y}^{(r)}(t)/1000$, où $F_{I,y}^{(r)}(t)$ désigne l'estimateur $\hat{F}_{I,y}(t)$ sur le $r^{\text{ème}}$ échantillon, $r = 1, \; \ldots \; , 1000$. Comme mesure de variabilité de $\hat{F}_{I,y}(t)$, nous avons utilisé

l'erreur quadratique moyenne relative (en pourcentage) de Monte Carlo

$$\text{RMSE}\{\hat{F}_{I,y}(t)\} = \frac{\sqrt{\text{MSE}\{\hat{F}_{I,y}(t)\}}}{F_{N,y}(t)} \times 100, \tag{22}$$

avec

$$\text{MSE}\{\hat{F}_{I,y}(t)\} = \frac{1}{1000}\sum_{r=1}^{1000}\{F_{I,y}^{(r)}(t) - F_{N,y}(t)\}^2. \tag{23}$$

			α				
			0.05	0.25	0.50	0.75	0.95
$\mathbf{x} = (1, x1, x2)$	REGI	RB	0.93	0.82	0.10	-0.11	-0.01
		RMSE	23.89	9.31	5.03	2.71	1.06
	REGI-P1	RB	1.00	0.59	0.02	-0.14	0.01
		RMSE	25.37	9.17	4.92	2.73	1.05
	REGI-P2	RB	0.24	0.83	0.18	-0.09	0.01
		RMSE	23.90	9.34	4.87	2.64	1.06
$\mathbf{x} = (1.x1)$	REGI	RB	-18.84	-12.90	-8.53	-4.56	-1.04
		RMSE	29.42	15.97	10.22	5.58	1.71
	REGI-P1	RB	0.52	0.40	0.10	-0.10	-0.01
		RMSE	27.45	9.82	5.28	2.84	1.10
	REGI-P2	RB	-19.15	-13.06	-8.86	-4.83	-1.08
		RMSE	30.36	16.36	10.57	5.86	1.73
	RHDI	RB	-29.25	-22.71	-16.22	-9.65	-2.44
		RMSE	37.71	24.73	17.41	10.42	3.04
	RHDI-P1	RB	0.59	0.19	0.07	-0.16	-0.02
		RMSE	34.15	11.73	5.99	3.08	1.19
	RHDI-P2	RB	-17.41	-13.06	-9.03	-4.98	-1.07
		RMSE	32.95	16.68	10.97	6.14	1.83

Tab. 1: Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour la fonction de répartition imputée

Les résultats obtenus sont donnés dans le tableau 1. Il est clair que lorsque le modèle d'imputation est correctement spécifié, l'imputation par la régression aléatoire conduit à une estimation non biaisée des paramètres, quels que soient les poids d'imputation utilisés. Lorsque le modèle d'imputation est mal spécifié, l'imputation par la régression aléatoire non pondérée (REGI) ou pondérée en se basant sur un modèle de non-réponse mal spécifié (REGI-P2) conduit à un biais important pour l'estimation des paramètres. En revanche, l'imputation par la régression aléatoire pondérée en se basant sur un modèle de non-réponse bien spécifié (REGI-P1) conduit à des estimateurs non biaisés. Les conclusions sont similaires pour l'imputation par hot-deck aléatoire.

5 Conclusion et travail futur

Dans cet article, nous montrons que l'imputation par le hot-deck aléatoire pondéré en se basant sur un modèle de non-réponse conduit à une estimation consistante de la

fonction de répartition. Dans le cas général d'une imputation par la régression aléatoire, on peut se demander s'il est possible de choisir des poids d'imputation conduisant au même résultat de consistance. Ce problème est actuellement à l'étude.

Le résultat obtenu est une convergence ponctuelle de la fonction de répartition. En pratique, on s'intéresse souvent à des quantiles, et obtenir une estimation consistante de ces paramètres nécessite un résultat de convergence uniforme. Ce problème est également à l'étude.

Références

[1] Breidt, F. J., and Opsomer, J. D. Local polynomial regresssion estimators in survey sampling. *Ann. Statist. 28*, 4 (2000), 1026–1053.

[2] Cao, W., Tsiatis, A. A., and Davidian, M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*, 3 (2009), 723–734.

[3] Chambers, R. L., and Dunstan, R. Estimating distribution functions from survey data. *Biometrika 73*, 3 (1986), 597–604.

[4] Chauvet, G., Haziza, D., and Deville, J.-C. On balanced random imputation in surveys. 459–471.

[5] Haziza, D. Imputation and inference in the presence of missing data. In *Handbook of Statistics, Sample Surveys : Theory Methods and Inference*, C. Rao and D. Pfeffermann, Eds. 1999, pp. 215–246.

[6] Haziza, D., and Rao, J. N. K. A nonresponse model approach to inference under imputation for missing survey data. 53–64.

[7] Kang, J. D. Y., and Schafer, J. L. Rejoinder : Demystifying double robustness : a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci. 22*, 4 (2007), 574–580.

[8] Kim, J. K., and Park, H. Imputation using response probability. *Canad. J. Statist. 34*, 1 (2006), 171–182.

[9] Kott, P. S. A note on handling nonresponse in sample surveys. *J. Amer. Statist. Assoc. 89*, 426 (1994), 693–696.

[10] Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc. 89*, 427 (1994), 846–866.

[11] Rubin, D. B. Inference and missing data. *Biometrika 63*, 3 (1976), 581–592. With comments by R. J. A. Little and a reply by the author.

[12] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc. 94*, 448 (1999), 1096–1146. With comments and a rejoinder by the authors.

[13] Tan, Z. A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc. 101*, 476 (2006), 1619–1637.

[14] Tan, Z. Comment : Understanding OR, PS and DR [mr2420458]. *Statist. Sci. 22*, 4 (2007), 560–568.

[15] Tan, Z. Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *Canad. J. Statist. 38*, 4 (2010), 609–632.

Annexe A

Démonstration du lemme 1

Notons E_p et V_p l'espérance et la variance sous le plan de sondage ; E_q et V_q l'espérance et la variance sous le mécanisme de réponse ; E_I et V_I l'espérance et la variance sous le mécanisme d'imputation. Pour montrer que $T_2 \to_{\mathbb{P}} 0$, il est équivalent de montrer que

$$\tilde{T}_2 = N^{-1} \sum_{i \in s} d_i (1 - r_i) \left\{ 1(y_i^* \leq t) - 1(\hat{y}_i \leq t) \right\} \quad \to_{\mathbb{P}} \quad 0.$$

Nous avons

$$E_I(|\tilde{T}_2|) \quad \leq \quad N^{-1} \sum_{i \in s} d_i (1 - r_i) \sum_{j \in s} \tilde{\omega}_j r_j \left| 1(E_j \leq t_{1,ij}) - 1(E_j \leq t_i) \right| \equiv \breve{T}_2,$$

où $t_i = \sigma_U^{-1} v_i^{-1/2} (t - \mathbf{x}_i^\top \mathbf{B}_U)$ et $t_{1,ij} = t_i + \sigma_U^{-1} \left(v_i^{-1/2} \mathbf{x}_i^\top - v_j^{-1/2} \mathbf{x}_j^\top \right) (\mathbf{B}_U - \hat{\mathbf{B}}_r)$. Soit $\eta > 0$, à spécifier ultérieurement, et

$$A_\eta = A_\eta^1 \cap A_\eta^2 \cap A_\eta^3$$

avec $A_\eta^1 = \{ \left| \hat{\mathbf{B}}_r - \mathbf{B}_U \right| \leq \eta \}$, $A_\eta^2 = \{ N^{-1} \left| \sum_{i \in s} \omega_j r_j - \sum_{i \in U} (1 - p_j) \right| \leq \eta \}$ and $A_\eta^3 = \{ \sup_{t \in \mathbb{R}} \left| \hat{F}_N(t) - \hat{F}_N(t) \right| \leq \eta \}$. Soit B_η le complémentaire de A_η. Sous les hypothèses $C1$-$C4$ et $C7$, nous avons

$$E_{pq} \left(\breve{T}_2 \; 1(B_\eta) \right) \to 0. \tag{24}$$

D'un autre côté, sous l'évènement A_η, l'hypothèse $C6$ implique qu'il existe une constante M_1 telle que pour tout (i, j) nous ayions $|t_{1,ij} - t_i| \leq M_1 \eta$. Après un peu de calcul, cela conduit à

$$\begin{aligned} E_{pq}(\breve{T}_2 \; 1(A_\eta)) \quad &\leq \quad \left[\sum_{j \in U} (1 - p_j) - N \; \eta \right]^{-1} \sum_{i \in U} \left[F_{N,E}(t_i + M_1 \eta) - F_{N,E}(t_i - M_1 \eta) \right] \\ &\leq \quad \frac{1}{1 - \eta - \kappa} \frac{1}{N} \sum_{i \in U} \left[F_{N,E}(t_i + M_1 \eta) - F_{N,E}(t_i - M_1 \eta) \right], \end{aligned}$$

où la dernière ligne est une conséquence de $C4$. Pour tout $\epsilon > 0$, l'hypothèse $C5$ implique qu'il existe $\eta > 0$ tel que pour tout $i \in U$

$$\left[F_{N,E}(t_i + M_1 \eta) - F_{N,E}(t_i - M_1 \eta) \right] \leq \epsilon.$$

En conséquence, nous obtenons

$$E_{pq} \left(\breve{T}_2 \; 1(A_\eta) \right) \to 0. \tag{25}$$

Avec l'équation (24), cela implique que $T_2 \to_{\mathbb{P}} 0$.

Annexe B

Démonstration du lemme 2

Pour montrer que $T_1 \to_{\mathbb{P}} 0$, il est équivalent de montrer que

$$\tilde{T}_1 = N^{-1} \sum_{i \in s} d_i (1 - r_i) \left\{ 1(\hat{y}_i \leq t) - 1(y_i \leq t) \right\} \quad \to_{\mathbb{P}} \quad 0.$$

Nous avons

$$\begin{aligned} E_I\left(\tilde{T}_1\right) &= N^{-1}\sum_{i\in s} d_i(1-r_i)\sum_{j\in s}\tilde{\omega}_j r_j\left\{1(y_j\leq t)-1(y_i\leq t)\right\} \\ &= U_1+U_2, \end{aligned}$$

avec

$$\begin{aligned} U_1 &= N^{-1}\left(\sum_{k\in s} d_k(1-p_k)\right)^{-1}\sum_{i\in s} d_i(1-r_i)\sum_{j\in s}\omega_j r_j\left\{1(y_j\leq t)-1(y_i\leq t)\right\}, \\ U_2 &= N^{-2}X\sum_{i\in s} d_i(1-r_i)\sum_{j\in s}\omega_j r_j\left\{1(y_j\leq t)-1(y_i\leq t)\right\}, \end{aligned}$$

et

$$X = \left(\frac{N}{\sum_{k\in s}\omega_k r_k}-\frac{N}{\sum_{k\in s} d_k(1-p_k)}\right).$$

Nous montrons tout d'abord que

$$E_{pqI}(\tilde{T}_1) \rightarrow 0. \tag{26}$$

Après un peu de calcul, nous obtenons que

$$E_q(U_1) = 0. \tag{27}$$

En utilisant des arguments similaires à ceux employés dans la démonstration précédente pour le terme $\breve{T}_2$, on montre que

$$E_{pq}(|U_2|) \rightarrow 0. \tag{28}$$

L'équation (26) découle des équations (27) et (28).

Nous montrons maintenant que

$$V_{pqI}(\tilde{T}_1) \rightarrow 0. \tag{29}$$

On a $V_{pqI}(\tilde{T}_1) = E_{pq}V_I(\tilde{T}_1)+V_{pq}E_I(\tilde{T}_1)$. Nous nous intéressons tout d'abord au premier terme.

$$\begin{aligned} V_I(\tilde{T}_1) &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)\sum_{j\in s}\tilde{\omega}_j r_j\left\{1(y_j\leq t)-\sum_{k\in s}\tilde{\omega}_k r_k 1(y_k\leq t)\right\}^2 \\ &\leq N^{-2}\sum_{i\in s} d_i(1-r_i). \end{aligned}$$

Les hypothèses $C2$, $C3$ et $C4$ impliquent que $V_I(\tilde{T}_1) = O(n^{-1})$, d'où

$$E_{pq}V_I(\tilde{T}_1) \rightarrow 0. \tag{30}$$

Nous considérons maintenant le terme

$$\begin{aligned} V_{pq}E_I(\tilde{T}_1) &= V_{pq}(U_1+U_2) \\ &= V_{pq}(U_1)+V_{pq}(U_2)+Cov_{pq}(U_1,U_2). \end{aligned}$$

En utilisant (27), nous avons $V_{pq}(U_1) = E_p V_q(U_1)$, et après un peu de calcul nous obtenons que $V_q(U_1) = O(n^{-1})$, d'où $V_{pq}(U_1) \to 0$. D'autre part, $V_{pq}(U_2) \leq E_{pq}(U_2^2)$, et en utilisant des arguments similaires à ceux employés dans la démonstration précédente pour le terme $\breve{T}_2$, on montre que $E_{pq}(U_2^2) \to 0$, d'où $V_{pq}(U_2) \to 0$. Par l'inégalité de Cauchy-Schwarz, on obtient $Cov_{pq}(U_1, U_2) \to 0$. On en déduit que

$$V_{pq} E_I(\tilde{T}_1) \quad \to \quad 0. \tag{31}$$

L'équation (29) découle des équations (30) et (31), et le résultat souhaité découle de l'inégalité de Bienaymé-Tchebyshev.